Bayesian Causal Inference - Session 2

Hierarchical Causal Inference, Sensitivity, and Nonparametrics

Arman Oganisian
Thomas J. & Alice M. Tisch Assistant Professor of Biostatistics

Department of Biostatistics
Brown University

5/29/2025

Outline

Session 1: Bayesian Crash Course

- Priors, Posteriors, Likelihoods.
- Posterior Computation.
- Shrinkage and Comparison with frequentist methods.

Session 2: Bayesian Causal Inference

- Hierarchical Causal Inference.
- Sensitivity Analysis.
- Bayesian ML methods.

Objectives

By the end of Session 2, I am hoping you will:

- Have an appreciation and high-level understanding of how hierarchical priors can be used to stabilize causal effect estimates in small samples.
- Have an understanding of how Bayesian methods can be used to express prior beliefs about causal assumptions.
- Have awareness of how Bayesian machine learning (ML) can be used to model complex covariate-outcome relationships, with automatic uncertainty quantification.

Hierarchical Causal Modeling

Notation

Consider a conditional average treatment effect (CATE) estimand

$$\Psi(v) = \mathrm{E}[Y^1 \mid V = v] - \mathrm{E}[Y^0 \mid V = v]$$

and define notation:

- V : discrete, pre-treatment covariate with K levels.
- L : single binary covariate.
- $A \in \{0, 1\}$: binary treatment indicator.
- Y^a: continuous, real-valued *potential* outcome under $A = a$.
- Y: continuous, real-valued *observed* outcome under $A = a$.

Full data, $D = \{v_i, l_i, a_i, y_i\}_{i=1}^n$.

We're keeping models/setting simple to illustrate the Bayesian features.

Identification

If $Y^a \perp A \mid L, V$ (and other usual causal assumptions), each term of $\Psi(v)$ is identified via the standardization formula

$$\Psi(v) = \int \big(\eta(1, l; \beta_v) - \eta(0, l; \beta_v)\big) f_L(l \mid v; \theta_v) dl$$

Within each stratum of V, we must model

- A regression function $\eta(a, l; \beta_v) = \mathrm{E}[Y \mid A = a, L = l; \beta_v]$
- A covariate distribution $f_L(l \mid v; \theta_v)$

Bayesian inference: requires obtaining posterior draws $f(\omega \mid D)$, where $\omega = (\beta_1, \theta_1, \ldots, \beta_K, \theta_K)$.

A posterior over ω induces a posterior over $\Psi(v)$.

One frequentist alternative, obtain $\hat{\omega}$ and plug-in.

A Simple Model

Suppose we specify a linear and additive regression model for each $V = v$:

$$\eta(a, l; \beta_v) = \beta_{0v} + \beta_{1v} a + \beta_{2v} l$$

then, evaluating the integral will show that

$$\Psi(v) = \beta_{1v}$$

- In applications, some strata of V may be sparsely populated.
- Should we collapse sparse strata? Should we throw out the data? How to choose which strata to toss/collapse? A garden of forking paths (Gelman, 2013).

A Simple Bayesian Model

For a Bayesian model, we must have a full probability model and a prior. Suppose

$$Y \mid A, L, V \sim N(\eta(a, l; \beta_v), \sigma_v^2)$$

with mean

$$\eta(a, l; \beta_v) = \beta_{0v} + \beta_{1v} a + \beta_{2v} l$$

where $\Psi(v) = \beta_{1v}$. Leads to likelihood,

$$\mathcal{L}(\beta_1, \beta_2, \dots, \beta_K \mid D) = \prod_{i=1}^{n} N(\eta(a_i, l_i; \beta_{v_i}), \sigma_{v_i}^2)$$

- We keep all strata - no matter how sparse, without collapsing.
- Unlike frequentist inference, (proper) Bayesian inference requires a full probability model of the outcome - first and second moment conditions are not enough.

A Hierarchical Prior

For $\Psi(v) = \beta_{1v}$, specify

$$\begin{aligned} \beta_{11}, \beta_{12}, \ldots, \beta_{1K} \mid \mu &\sim N(\mu, \phi^2) \\ \mu &\sim N(0, \tau^2) \end{aligned}$$

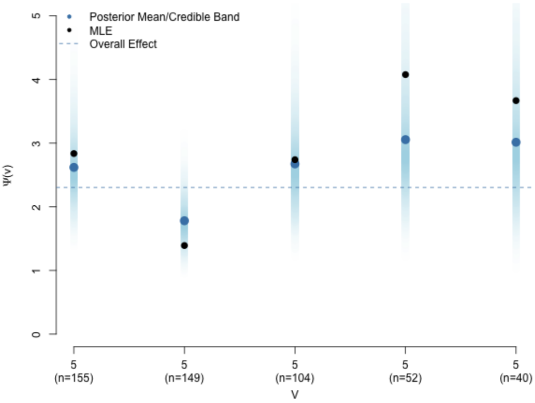
with a flat prior $f(\tau, \phi) \propto 1$.

A priori, β_{1v} are conditionally independent given μ, but marginally *dependent*

$$f(\beta_{11}, \beta_{12}, \ldots, \beta_{1K}) = \int \prod_{k=}^{K} f(\beta_{1k} \mid \mu) f(\mu) d\mu$$

- For sparse stratum, $V = k$, β_{1k} is shrunk towards the overall mean effect.
- In the absence of data, reasonable to think that causal effect in stratum k is similar to the average of the effect across the other strata.

Posterior Estimates

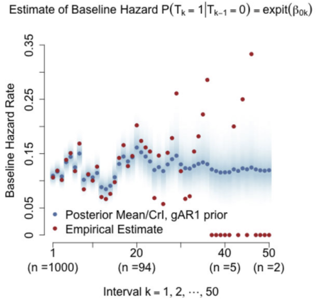


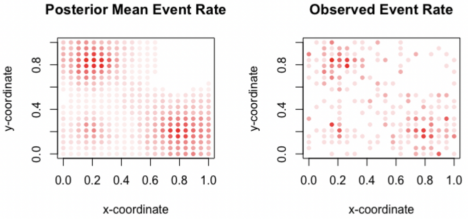
See Oganisian & Roy (2021) Sec. 3.1

Hierarchical Shrinkage - Other structures

In the previous example, the shrinkage structure was symmetric, but other shrinkage structures can be encoded in a similar way:

- Autoregressive shrinkage: useful for modeling temporal processes[a b] or "smooth" treatments such as dose level.[c]
- Spatial shrinkage: useful for modeling effects on outcomes measured across space.[d]





[a] See Oganisian et al. (2024) for recurrent event outcome example.

[b] See Han & Oganisian (2025) for survival outcome example.

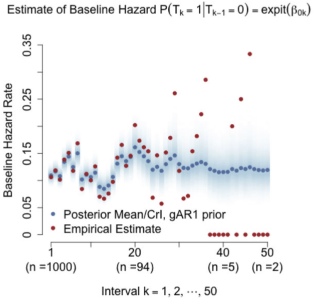
[c] Oganisian & Roy (2021) Sec. 3.1

[d] See Wikle & Zigler (2024) for example with air pollution.

Non-Bayesian Shrinkage

Although priors are subjective, non-Bayesian analyses smooth subjectively as well - via ad-hoc and trial-and-error methods:

- Young et al (2020): intercept set to be a second-order polynomial function of k "after several bootstrap samples for the construction of confidence intervals failed to converge under the more flexible model."
- Hernán et al. (2000): "We cannot estimate a separate intercepts for each month k. Rather, we need to 'borrow strength' from subjects starting zidovudine in months other than k to estimate $[\beta_{0k}]$. This can be accomplished by assuming that $[\beta_{0k}]$ is constant in windows of, say, 3 months."
- Dodd et al. (2019): "Taking into account the frequency and duration of follow-up information in this analysis with the potential for covariate information to be updated on a daily basis, it seemed sensible to use fortnightly intervals."



What have we learned?

A common misconception about Bayesian inference is that priors can be used to “cherry-pick” preferred parameter values. However,

- Rather than specifying values, priors can be used to specify dependence structures.
- These dependence structures can be tailored to induce smoothness.
- We also do smoothing during frequentist inference, but often in non-transparent and ad-hoc ways.
- Even with complex shrinkage, interval estimation for causal effects are automatic.

Sensitivity Analysis

Violations of Causal Assumptions

"Nothing is wrong with making assumptions; causal inference is impossible without making assumptions, and they are the strands that link statistics to science. It is the scientific quality of those assumptions, not their existence, that is critical." - Rubin (1986).

Causal inference - Bayesian or frequentist - requires that treatment be exchangeable conditional on observed covariates,

$$Y^a \perp A \mid L$$

- In many settings, there may be common cause(s) of treatment and potential outcomes that are unmeasured.
- In these settings $Y^a \not\perp A \mid L$.
- This assumption is not testable with/informed by data.
- However, prior beliefs in this assumption can be baked into a Bayesian analysis.

Bayesian Sensitivity Analysis

Consider a setting where we wish to compute

$$\Psi = P(Y^1 = 1) - P(Y^0 = 1)$$

However, while

$$Y^a \not\perp A \mid L$$

we think

$$Y^a \perp A \mid L, U$$

where

- A: binary treatment (anthracyclene chemotherapy).
- L: single *measured* confounder.
- Y: binary outcome - event is beneficial (e.g. 1-yr remission).
- U: single *unmeasured* confounder - e.g. ejection fraction (EF).

Observed data is $D = \{a_i, l_i, y_i\}_{i=1}^n$

Bayesian Sensitivity Analysis

If we have U observed, then standardization formula says

$$\Psi = \int\int \big(\eta(1, l, u; \beta) - \eta(0, l, u; \beta)\big) f(l, u; \theta) dl du$$

where $\eta(a, l, u; \beta) = P(Y = 1 \mid A = a, L = l, U = u; \beta)$

- We could specify some outcome regression model, e.g.

$$\eta(a, l, u; \beta) = g^{-1}(\beta_0 + \beta_1 a + \beta_2 l + \xi_1 u)$$

  with parameters $\beta = (\beta_0, \beta_1, \beta_2, \xi_1)$

- A bivariate density model $f(l, u; \theta)$.
- Estimates of β and θ yield an estimate of Ψ.
- But how could we ever estimate these if U is not observed?

Bayesian Sensitivity Analysis

From a Bayesian perspective, there is no difficulty. The procedure is always the same: find a distribution over quantities you want to know, conditional on the quantities you do know.

$$f(\beta, \theta \mid D) = \int f(\beta, \theta, \boldsymbol{u} \mid D) d\boldsymbol{u}$$

where $\boldsymbol{u} = (u_1, u_2, \ldots, u_n)$.

- The unmeasured confounder for each subject is treated as missing data - just another unknown, like β and θ
- Procedure: simulate from the joint posterior $f(\beta, \theta, \boldsymbol{u} \mid D)$, retain the draws of (β, θ) for inference.

Bayesian Sensitivity Analysis

It can be shown that

$$
\begin{aligned}
f(\beta, \theta, \boldsymbol{u} \mid D) \propto \prod_{i=1}^{n} & P(A = a_i \mid l_i, u_i; \gamma) f(Y = y_i \mid a_i, l_i, u_i; \beta) f(l_i, u_i; \theta) \\
& \times f(\beta, \gamma, \theta)
\end{aligned}
$$

where

- $P(A = 1 \mid l, u; \gamma)$ is a propensity score model. e.g.

  $$P(A = 1 \mid l, u; \gamma) = g^{-1}(\gamma_0 + \gamma_1 l + \xi_2 u)$$

  with parameters $\gamma = (\gamma_0, \gamma_1, \xi_2)$
- It is usually assumed that $U \perp L$ so that $f(l_i, u_i; \theta) = f(l_i; \theta) N(u_i; 0, 1)$.
- The unnormalized posterior above can be specified in `Stan` to obtain draws of $(\beta^{(m)}, \theta^{(m)})$.

Bayesian Sensitivity Analysis

$$\eta(a, l, u; \beta) = g^{-1}(\beta_0 + \beta_1 a + \beta_2 l + \xi_1 u)$$
$$P(A = 1 \mid l, u; \gamma) = g^{-1}(\gamma_0 + \gamma_1 l + \xi_2 u)$$

The quantities ξ_1 and ξ_2, are sensitivity parameters

- They encode departures from (violations of) conditional exchangeability.
- Priors on them encode our prior beliefs about the direction/magnitude the violation.
- E.g. $\xi_1, \xi_2 \sim \delta_0$ is a strong prior belief in no unmeasured confounding.
- There is no information in the observed data about them. We will need to compensate with informative priors.

Bayesian Sensitivity Analysis

Priors on ξ_1 and ξ_2 encode our prior beliefs about the direction/magnitude the violation. For example,

- Strong prior belief in no unmeasured confounding:

$$\xi_2 \sim \delta_0$$

- Belief in unmeasured confounding, but symmetric uncertainty in direction:

$$\xi_2 \sim N(0, 1)$$

- Belief in unmeasured confounding, with treated patients having higher U:

$$\xi_2 \sim Gam(1, 1)$$

- Strong prior belief in particular values ξ_2^* - perhaps by elicitation or from previous literature

$$\xi_2 \sim \delta_{\xi_2^*}$$

An Example

See example R file.

With our data, D, we want to estimate $\Psi = \mathrm{E}[Y^1 - Y^0]$.

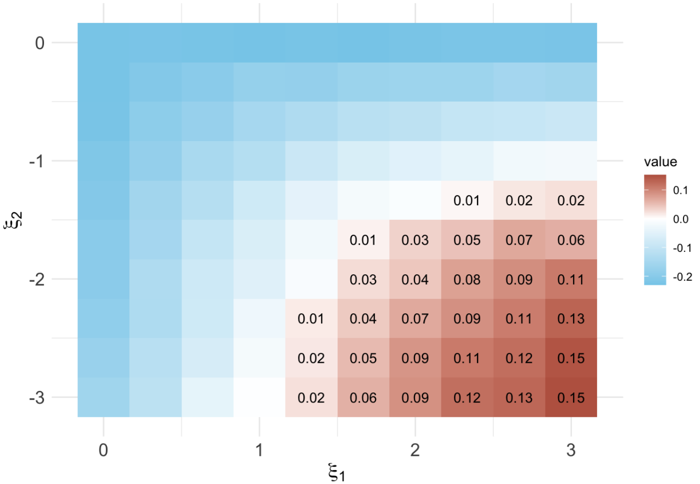
Under $Y^a \perp A \mid L$, Bayesian standardization yields posterior mean of

$$-.31, 95\% \text{ CrI:}[-.39, -.23]$$

- It seems treatment $A = 1$ lower remission rate.
- But what if patients with higher EF (U) were more likely to get $A = 1$ ($\xi_2 > 0$), and less likely to have remission ($\xi_1 < 0$)?
- Then we cannot determine whether lower remission rate was due to $A = 1$.
- Question: how large ξ_1 and ξ_2 need to be in their respective directions before the upper limit of the CrI exceeds 0 (i.e. posterior puts substantial probability around no effect)?

An Example

Upper limit of 95% CrI across priors. At $\xi_1 = \xi_2 = 0$, -.31 [-.39,-.23]



If we think a unit increase in U increases odds of event and treatment by $\approx e^1 \approx 2.7$, then $\Psi = 0$ would be in our 95% credible interval.

What have we learned?

In Bayesian inference, there is no distinction between “parameters” and “data.” This makes it uniquely suited for sensitivity analyses:

- Everything is a random variable - some are known at the time of analysis (data), others are unknown (parameters).
- A unmeasured confounder is simply another unknown - can turn the same Bayesian crank. No special methods are needed.
- We have different degrees of uncertainty about identification assumptions. Priors allow a direct way of encoding this uncertainty.
- Posterior reflects uncertainty at all levels - sampling variability, uncertainty about unknown parameters, uncertainty about identification assumptions.

A Glimpse of Bayesian ML for Causal Inference

Motivation for Flexible Inference

The standardization formula

$$\Psi = \int \big(\eta(1,l) - \eta(0,l)\big) f_L(l) dl$$

where $\eta(a,l) = E[Y \mid A = a, L = l]$.

- Previously we looked at parametric models for $\eta(a,l)$ and $f_L(l)$.
- But these cannot capture complex outcome-covariate relationships.
- We want models that allow for complicated functional forms if warranted, but shrink back to a simple form.
- For simplicity, let's just call it $\eta(x;\beta)$, where $x = (a,l)$.
- We'll ignore $f_L(l)$ for now, but imagine using the empirical distribution for this.

Bayesian Additive Regression Trees

Bayesian Additive Regression Trees

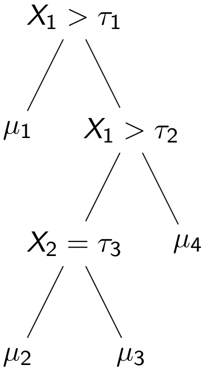
$$\eta(x) = \sum_{j=1}^{m} g\left(x; T_j, M_j\right)$$

- m trees, T_j for $j = 1, 2 \ldots, m$: each consisting of a set of terminal nodes, M_j, a set of splitting thresholds, $\bar{\tau}_j$, and some depth d.
- Each terminal node in M_j has associated parameter η_k.
- $g(\cdot)$ maps x to some $\mu_k \in M_j$ of tree T_j via conditional logic statements.
- The original BART assumes $\epsilon \sim N(0, \sigma^2)$.

Chipman et al. (2010)

Parameters of a Tree Structure

Tree T_j with terminal node parameters $M_j = \{\mu_1, \mu_2, \mu_3, \mu_4\}$



Unknown parameters that determine tree structure are $\bar{\tau}_j = (\tau_1, \tau_2, \tau_3)$ and terminal node parameters M_j. The tree depth (number of layers), d, is also unknown. The choice of variable to split on at each layer is also unknown.

BART Priors

BART a prior *process* that generates random tree functions via priors on all of the tree unknowns.

- At a given layer in the tree, choose a variable in X to split on (uniform prior).
- Choose a splitting threshold uniformly from each unique observed.
- Choose whether to grow the tree for one additional later or to stop according to some Bernoulli probability.
- Generate terminal node parameters, η, from some prior distribution.

m trees are grown in this way. For a given subject, an estimate of $\eta(x) = E[Y \mid X = x]$ is formed by summing the terminal parameters of each tree.

Prior over Tree Depth

Probability that node at depth d is non-terminal

$$\frac{\alpha}{(1+d)^{\beta}}$$

For $\alpha \in (0,1)$ and $\beta \geq 0$

- For $\beta > 0$ Probability of depth d being non-terminal decreases with d. That is, if the tree is already large it is less likely to grow.
- This regularizes the tree by shrinking towards “shallow” trees.
- For $\beta > 1$ this shrinkage is more aggressive.

Prior over Tree Depth

Probability that node at depth $d = 0, 1, 2, 3, \dots$ is non-terminal

$$\frac{\alpha}{(1+d)^{\beta}}$$

For $\alpha \in (0,1)$ and $\beta \geq 0$

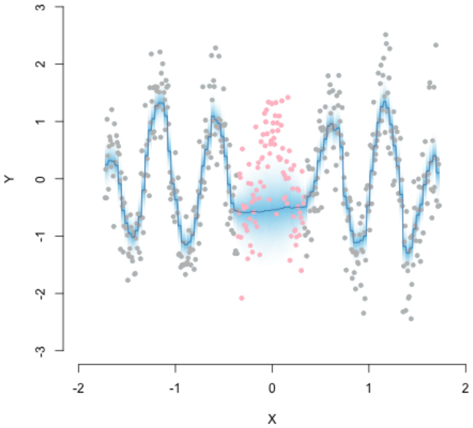
- For $\beta > 0$ Probability of depth d being non-terminal decreases with d. That is, if the tree is already large it is less likely to grow.
- This regularizes the tree by shrinking towards "shallow" trees.
- For $\beta > 1$ this shrinkage is more aggressive.
- Default values of $\alpha = .95$ and $\beta = 2$.

Putting it all together, we say the function $\eta(x)$ follows a BART prior

$$\eta \sim BART$$

Bayesian Additive Regression Trees

As implemented with `bayestree` package in `R`.



Remarks

Compare with ad hoc approaches in ML such as random forest:

- Splitting determined by various “purity” criterion - of which there are many possible choices.
- Regularization is not built in - requires “pruning” steps.
- Difficult to get valid uncertainty estimates.

In contrast, BART

- Determines splits probabilistically based on increases in unnormalized posterior density evaluations.
- Regularization is built-in via a prior on the tree depth - can grow the tree large if data are complex, but shrink to a “stump” if complexity is not warranted.
- BART is a valid prior process which leads to posterior draws of the trees - uncertainty in predictions reflected across draws.

A major benefit: default hyperpriors that work well off the shelf.